Information Systems

Dr Sherin El Gokhy

Topics : Data Science and Big Data Analytics

Introduction to Big Data Analytics + Data Analytics Lifecycle	Review of Basic Data Analytic Methods Using R	Advanced Analytics – Theory and Methods	Advanced Analytics - Technology and Tools	The Endgame, or Putting it All Together + Final Lab on Big Data Analytics
Big Data Overview State of the Practice in Analytics The Data Scientist Big Data Analytics in Industry Verticals Data Analytics Lifecycle	Using R to Look at Data - Introduction to R Analyzing and Exploring the Data Statistics for Model Building and Evaluation	K-means Clustering Association Rules Linear Regression Logistic Regression Naive Bayesian Classifier Decision Trees Time Series Analysis Text Analysis	Analytics for Unstructured Data (MapReduce and Hadoop) The Hadoop Ecosystem In-database Analytics – SQL Essentials Advanced SQL and MADlib for In-database Analytics	Operationalizing an Analytics Project Creating the Final Deliverables Data Visualization Techniques + Final Lab – Application of the Data Analytics Lifecycle to a Big Data Analytics Challenge

- **“Big Data”** is a popular term which refers to the exponential growth and availability of data, both **structured** and **unstructured**
- In **2001**, industry analyst **Doug Laney** attributed the “**three Vs”** to describe the definition of big data
  - Volume
  - Velocity
  - Variety

Volume

There has been a large increase of data volume. There are multiple reasons for this..

- All of the transactional data that has added up over the years
- Streaming data from social media
- Machine to machine data increase

Initially, storage was a big concern but with costs of storage dropping, it is not as big of a threat as things like analytics.

So, Data Analytical is the main topic that is concerned in this course.

Velocity

Data is being streamed at huge speeds and needs to be dealt with in a timely manner. Some examples are…

- Social Media
- Mobile Devices

The biggest challenge is how to react fast enough to the massive amount of data that is being flew rapidly

Variety

There are many different types of data

- Structured Data
- Numeric Data
- Application Data
- Unstructured Documents
- Email
- Audio & Video
- Financial Transactions

Managing all the different formats is an issue many organizations have to battle

4 V's

- Volume (size)
- Velocity (rapidly streaming)
- Variety (many forms)
- Veracity: Uncertainty of data

refers to the trustworthiness of the data. With many forms of big data, quality and accuracy are less controllable (just think of Twitter posts with hash tags

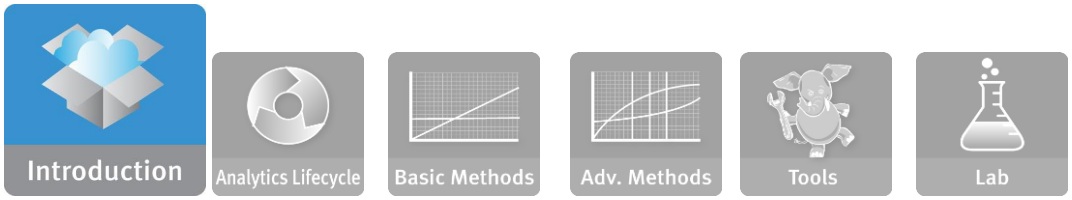
5 V's…. Value of data is added

Well and good for access or useless data (Business value of data)

Introduction

Big Data Analytics

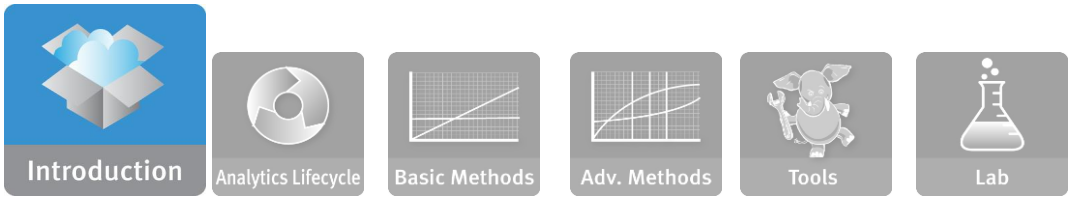
Module 1 – Introduction to Big Data Analytics



Module 1: Introduction to Big Data Analytics

Upon completion of this module, you should be able to:

- Define big data
- Identify four business drivers for advanced analytics
- Distinguish the techniques for Business Intelligence from those of Data Science
- Describe the role of the Data Scientist within the new big data ecosystem
- Cite at least three illustrative examples of big data opportunities



Module 1: Introduction to Big Data Analytics

Big Data Overview

During this part the following topics are covered:

- Definition of big data
- Big data characteristics and considerations
- Unstructured data supporting big data analytics
- Analyst perspective on Data Repositories (the evolution of data repositories)

What is *Big Data*?

What makes data, "*Big" Data*?

Big Data Definition

- ***"Big Data" is data whose scale, distribution, diversity, and/or timeliness require the use of new technical architectures and analytics to enable insights that unlock new sources of business value.***
  - Requires new data architectures, analytic sandboxes
  - New tools/ technologies to store, manage and realize the business benefit of these large data sets
  - New analytical methods
  - Integrating multiple skills into new role of data scientist
- Organizations are deriving business benefit from analyzing ever larger and more complex data sets **that increasingly require real-time or near-real time capabilities**
- Big Data is not just a scientific term. It has a business value.

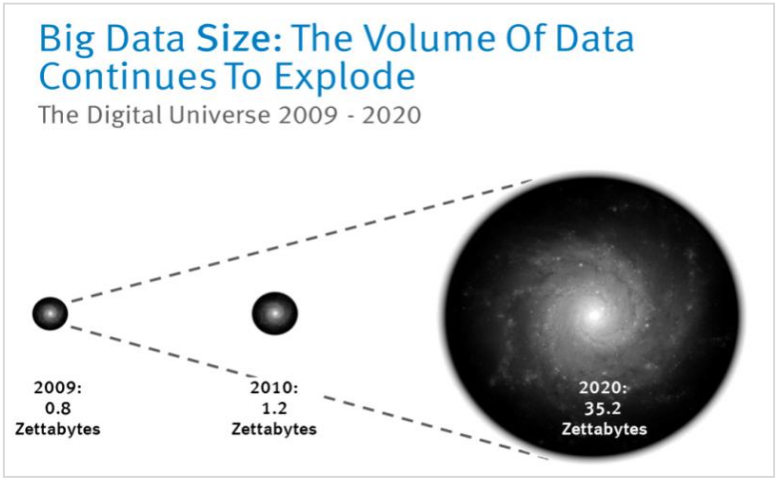
Source: McKinsey May 2011 article *Big Data: The next frontier for innovation, competition, and productivity*

Key Characteristics of Big Data

1. Data Volume

- 44x increase from 2009 to 2020
  (0.8 zettabytes to 35.2zb)

Highly rate of growth (very accelerating)



2. Processing Complexity

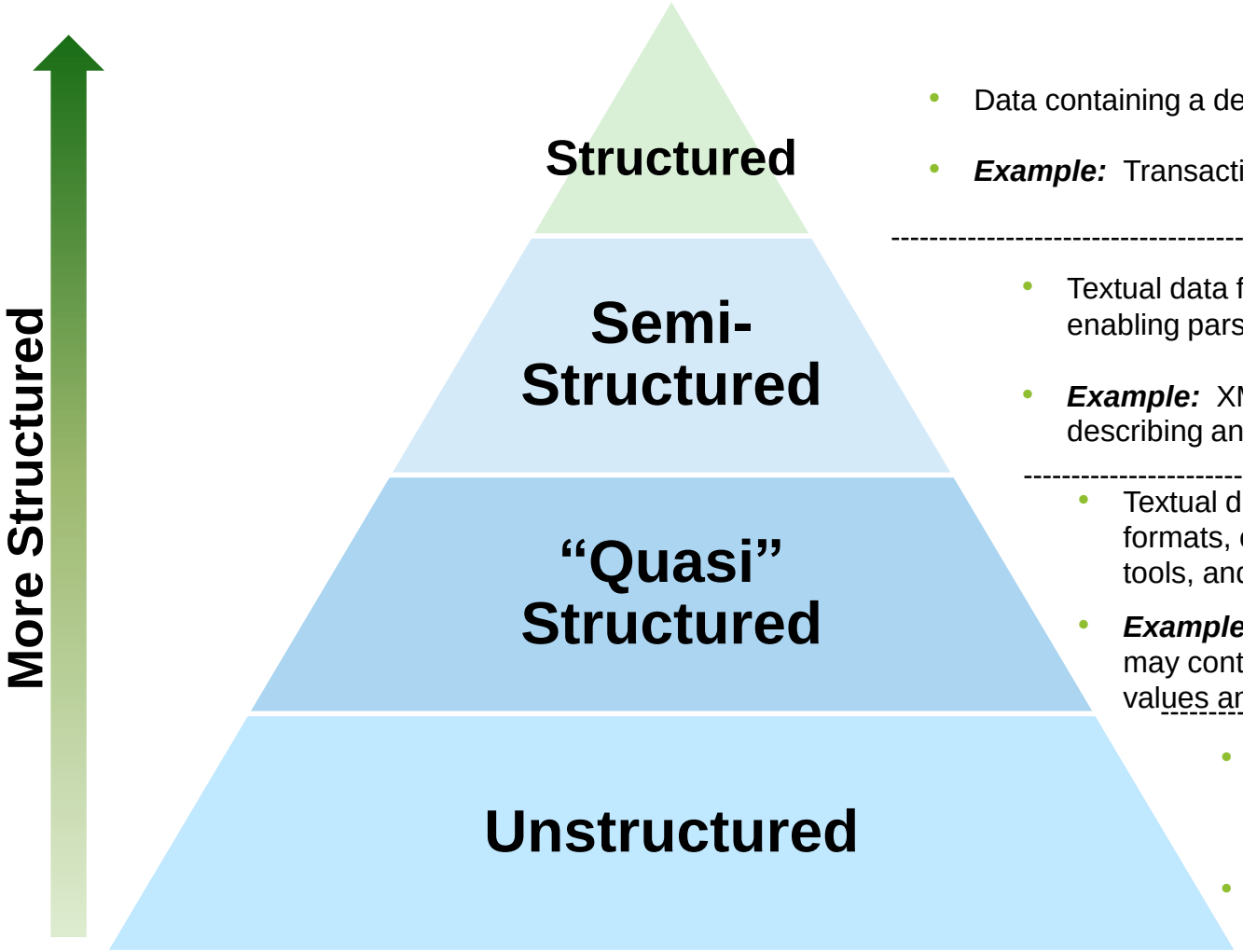
- Changing data structures
- Use cases requires additional transformations and different analytical techniques
- The preferred approach for processing big data is in **parallel computing environments and Massively Parallel Processing** , which enable simultaneous, parallel loading and analysis of data.

3. Data Structure

- Greater variety of data structures to mine and analyze
- Most of the big data is unstructured or semi-structured in nature, **which requires different techniques and tools to process and analyze.**

Big Data Characteristics: Data Structures

Data Growth is Increasingly Unstructured



- Data containing a defined data type, format, structure
- ***Example:*** Transaction data and OLAP

- Textual data files with a discernable pattern, enabling parsing
- ***Example:*** XML data files that are self describing and defined by an xml schema

---

- Textual data with unorganized data formats, can be formatted with effort, tools, and time
- ***Example:*** Web clickstream data that may contain some inconsistencies in data values and formats

---

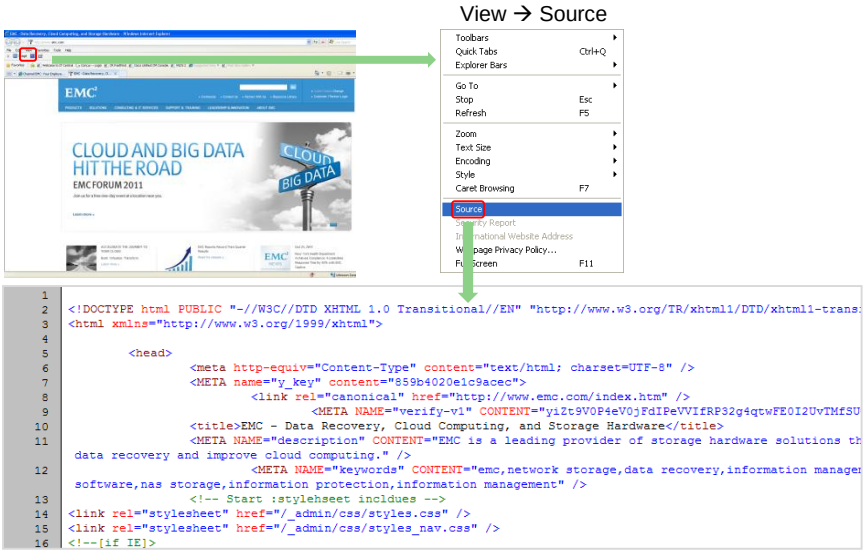
- Data that has no inherent structure and is usually stored as different types of files.
- ***Example:*** Text documents, PDFs, images and video

Four Main Types of Data Structures

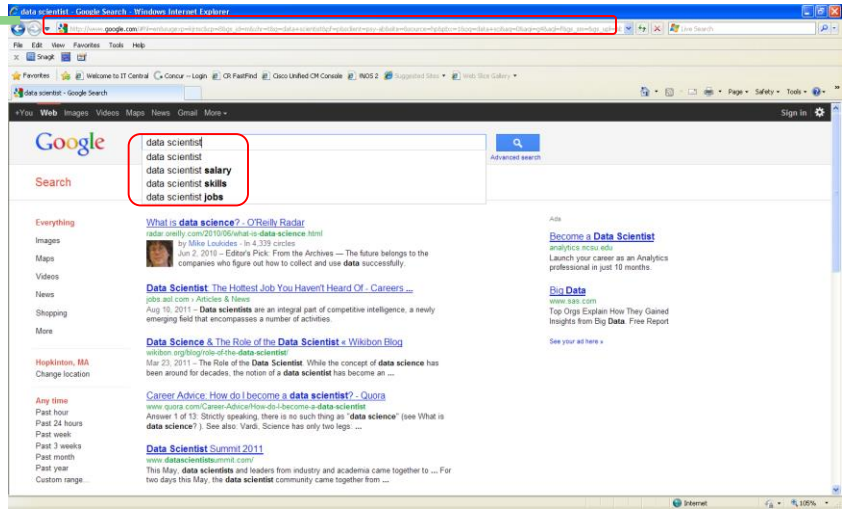
Structured Data

SUMMER FOOD SERVICE PROGRAM 1]				
(Data as of August 01, 2011)				
Fiscal Year	Number of Sites	Peak (July) Participation	Meals Served	Total Federal Expenditures 2]
	----------Thousands----------		--Mil.--	---Million $---
1969	1.2	99	2.2	0.3
1970	1.9	227	8.2	1.8
1971	3.2	569	29.0	8.2
1972	6.5	1,080	73.5	21.9
1973	11.2	1,437	65.4	26.6
1974	10.6	1,403	63.6	33.6
1975	12.0	1,785	84.3	50.3
1976	16.0	2,453	104.8	73.4
TQ 3]	22.4	3,455	198.0	88.9
1977	23.7	2,791	170.4	114.4
1978	22.4	2,333	120.3	100.3
1979	23.0	2,126	121.8	108.6
1980	21.6	1,922	108.2	110.1

Semi-Structured Data



Quasi-Structured Data



http://www.google.com/#hl=en&sugexp=kjrmc&cp=8&gs_id=2m&xhr=t&q=data+scientist&pq=big+data&pf=p&sclient=psyb&source=hp&pbx=1&oq=data+sci&aq=0&aqi=g4&aql=f&gs_sm=&gs_upl=&bav=on.2,or.r_gc.r_pw.,cf.osb&fp=d566e0fbd09c8604&biw=1382&bih=651

Unstructured Data

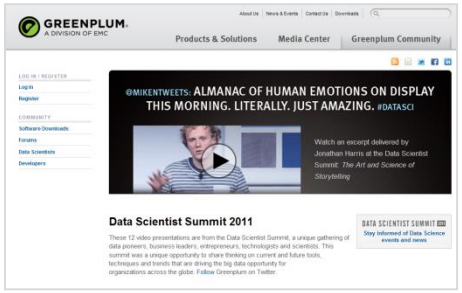
The Red Wheelbarrow, by William Carlos Williams

so much depends
upon

a red wheel
barrow

glazed with rain
water

beside the white
chickens.



Data Repositories, An Analyst Perspective

Data Islands
"Spreadmarts"

Isolated data marts

- Spreadsheets and low-volume DB's for record/ keeping
- Analyst dependent on data extracts

Data Warehouses

Centralized data containers in a purpose-built space

- Supports BI and reporting, but restricts robust analyses or data exploration
- Analyst dependent on IT & DBAs for data access and schema changes
- Analysts must spend significant time to get extracts from multiple sources

Analytic Sandbox

Data assets gathered from multiple sources and technologies for analysis

- Enables high performance analytics using in-db processing
- Reduces costs associated with data replication into "shadow" file systems
- *"Analyst-owned" rather than "DBA owned"*
- More robust analyses

Introduction to Big Data Analytics: Mini-Case Study

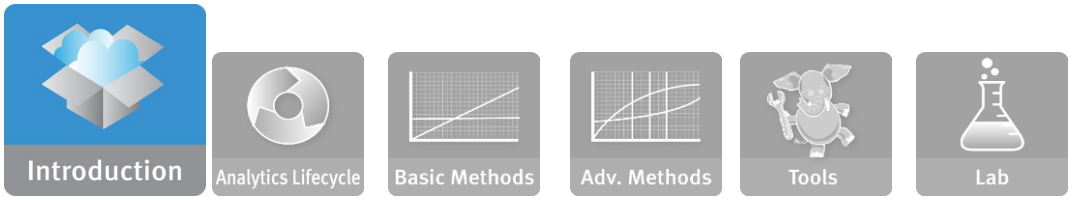
Yoyodyne Bank Scenario

- Evolving from small community bank to a global bank
- Needs to move away from its inheritance mainframes to an environment that supports more robust analytics
- Growing through mergers and acquisitions
- Subject to many new regulatory requirements
- Increasing customer base and increased product offerings

Your Thoughts?

Discussion Questions

1. Discuss how the bank's data would change under these circumstances.
2. How are their needs changing with these business changes?
3. What do you need to consider from an analyst point of view? What are some things to consider implementing as the bank grows?



Module 1: Introduction to Big Data Analytics

Summary

During this part the following topics were covered:

- Definition of big data
- Big data characteristics and considerations
- Unstructured data fueling big data analytics
- Analyst perspective on Data Repositories

Module 1: Introduction to Big Data Analytics

State of the Practice in Analytics

During this part the following topics are covered:

- Business drivers for analytics
- Current analytical architecture
- Business intelligence vs. data science
- Drivers of big data and new big data ecosystem

Business Drivers

***Business* Drivers: People, *knowledge*, and *conditions* that initiate and support *activities* for which the *business* was *designed*.**

Current Business Problems Provide Opportunities for Organizations to Become More Analytical & Data Driven

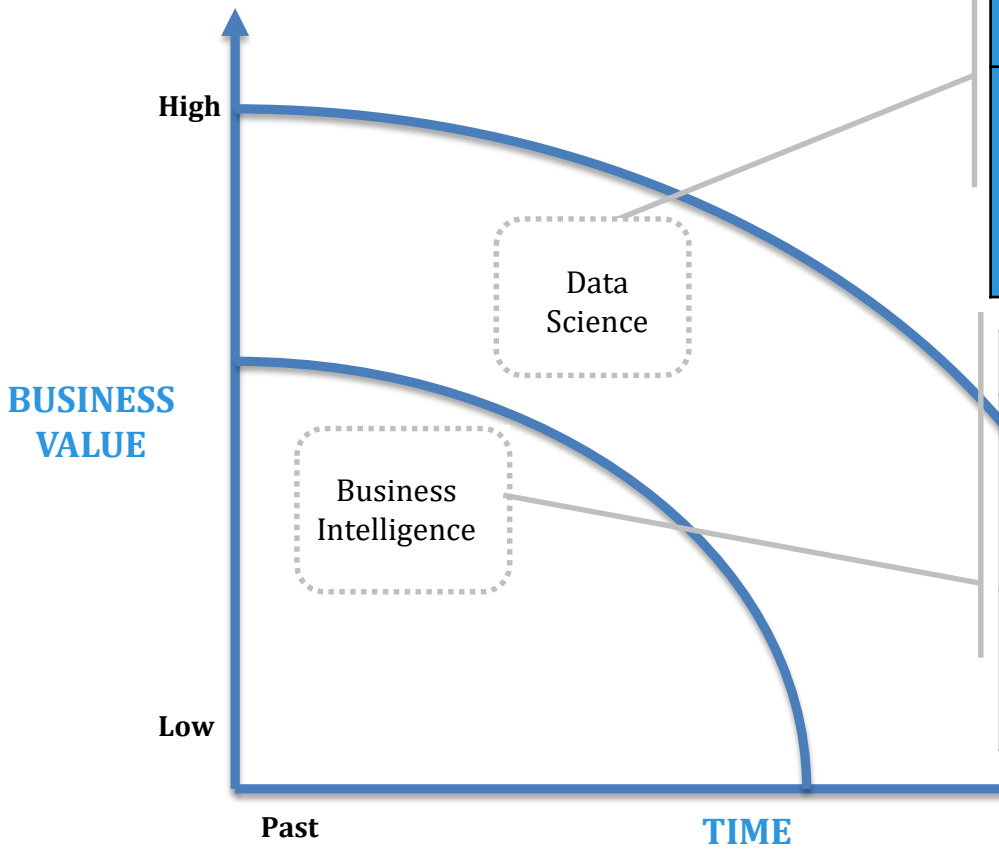
Business Drivers for Analytics

Here are 4 examples of common business problems that organizations contend with today, where they have an opportunity to do advanced analytics to create competitive advantage. Rather than doing standard reporting on these areas

	Driver	Examples
1	Desire to optimize business operations and derive more values from these typical tasks	Sales, pricing, profitability, efficiency
2	Desire to identify business risk to reduce it	Customer churn, fraud, default
3	Predict new business opportunities	Upsell, cross-sell, best new customer prospects
4	Obey laws or regulatory requirements	Anti-Money Laundering, Fair Lending

Analytical Approaches for Meeting Business Drivers

Business Intelligence vs. Data Science



Predictive Analytics & Data Mining (Data Science)	
Typical Techniques & Data Types	• Optimization, predictive modeling, statistical analysis, machine learning techniques as Naïve Bayes or regression • Structured/unstructured data, many types of sources, very large data sets
Common Questions	• What if…..? • What's the optimal scenario for our business ? • What will happen next? What if these trends continue? Why is this happening? **Open ended questions**

Business Intelligence	
Typical Techniques & Data Types	• Standard and ad hoc reporting, dashboards that provides KPIs, alerts, queries, details on demand • Structured data, traditional sources, manageable data sets
Common Questions	• What happened last quarter? • How many did we sell? • Where is the problem? In which situations?

A Typical Analytical Architecture





1 Data Sources

2 "Spread Marts"

Departmental Warehouse

Departmental Warehouse

Non-Agile Models

Enterprise Applications

3 Prioritized Operational Processes

Reporting

4

Analytics

Static schemas over time

Non-Prioritized Data Provisioning

EDW

Implications of Typical Architecture for Data Science

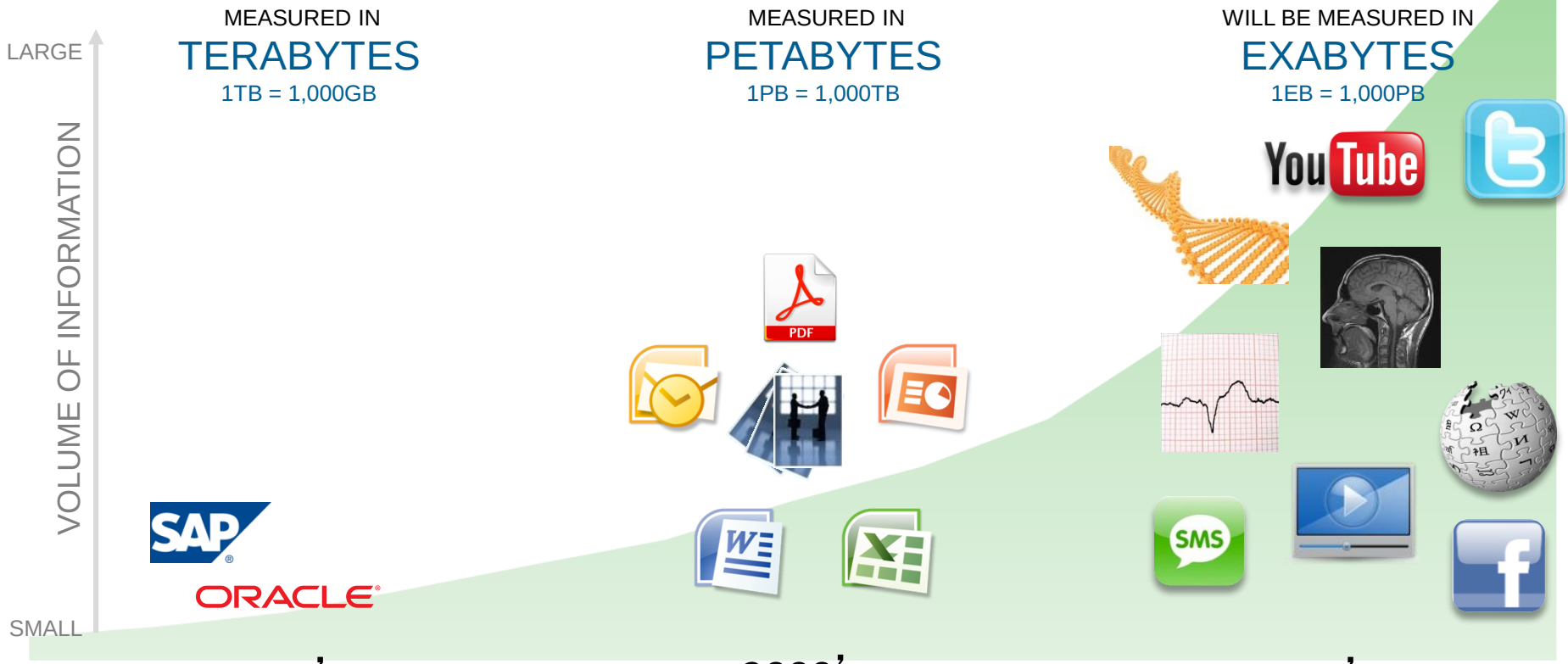
- High-value data analytics is hard to reach and leverage
- Predictive analytics & data mining activities are last in line for data
  - Queued after prioritized operational processes
- Data is moving in batches from EDW to local analytical tools
  - In-memory analytics (such as R, SAS, SPSS, Excel)
  - Sampling can skew model accuracy
- Isolated, *ad hoc* analytic projects, rather than centrally-managed of analytics
  - Frequently, not aligned with corporate business goals

Slow
"time-to-insight"
&
reduced
business impact

Opportunities for a New Approach to Analytics

New Applications Driving Data Volume

The Big Data trend is generating an enormous amount of information that requires advanced analytics and new market players to take advantage of it.



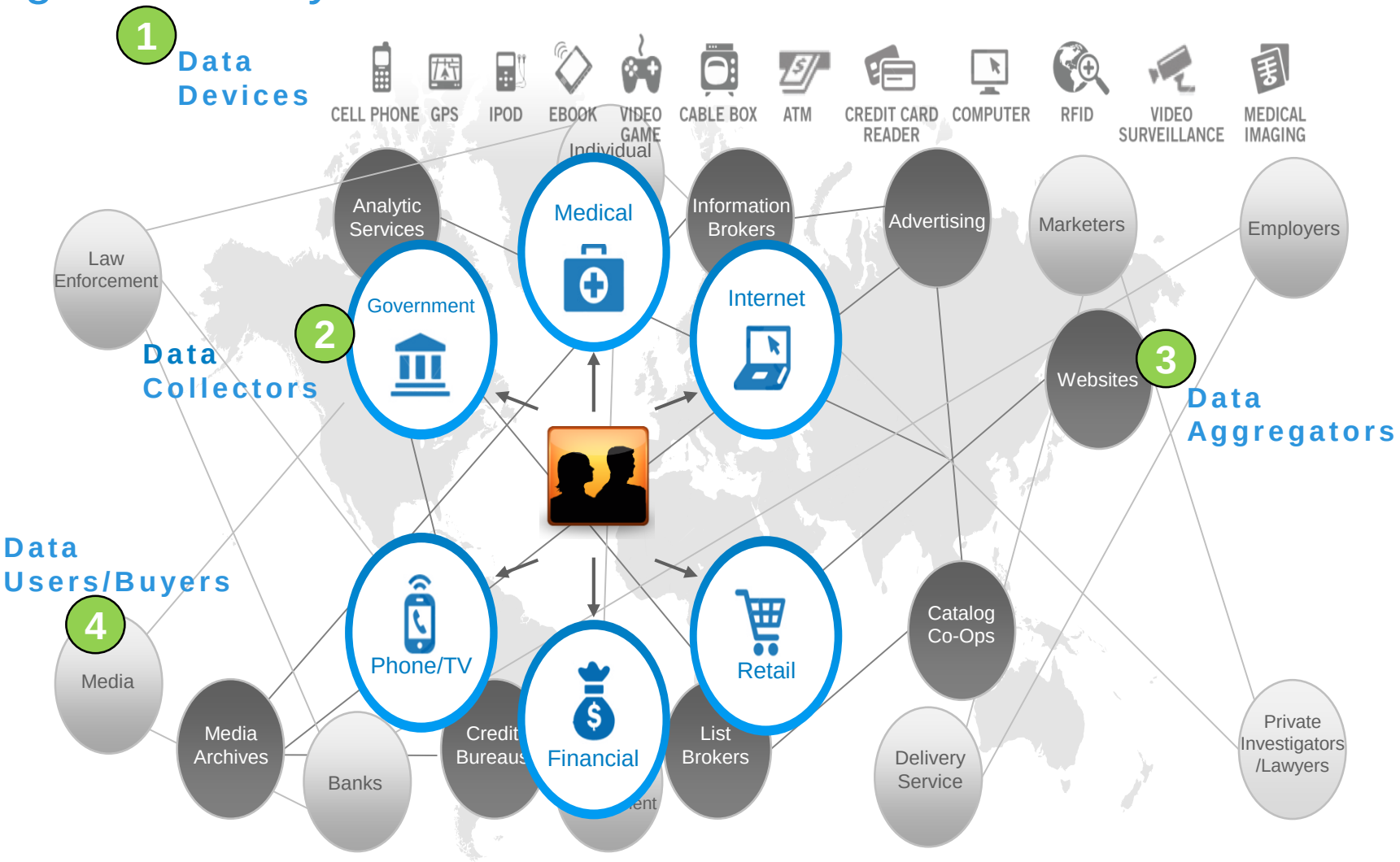
1990's (RDBMS & DATA WAREHOUSE)

2000's (CONTENT & DIGITAL ASSET MANAGEMENT)

2010's (NO-SQL & KEY/VALUE)

Opportunities for a New Approach to Analytics

Big Data Ecosystem



Opportunities for a New Approach to Analytics (Continued)
Big Data Ecosystem

Key Concepts:

A) Significant opportunities exist to extract value from Big Data

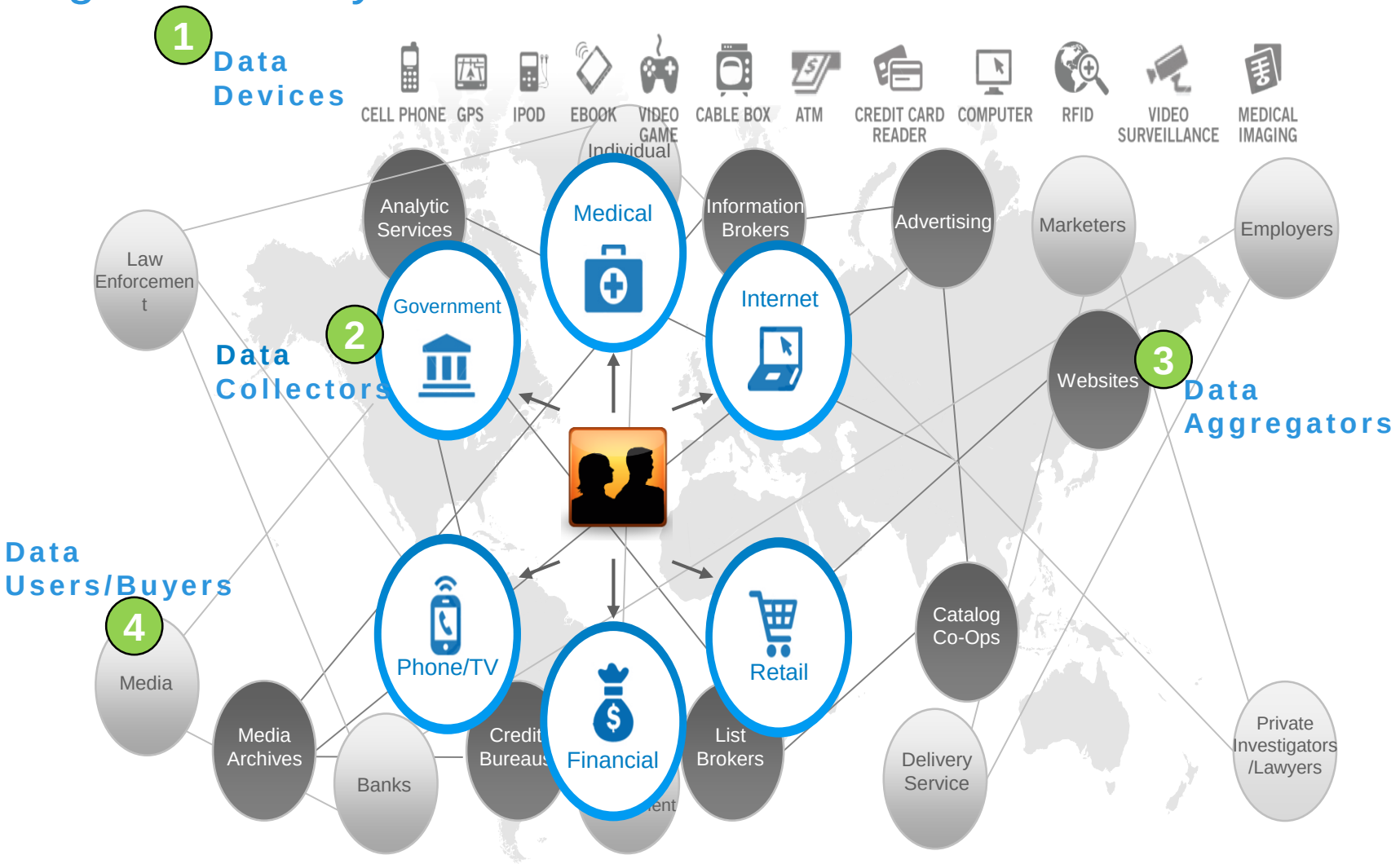
B) Entities are emerging throughout the new Big Data ecosystem to capitalize on these opportunities – from

1. ***Data Devices,***
2. ***Data Collectors,***
3. ***Data Aggregators,***
4. ***Data Users / Buyers***

C) To accomplish this, these players will need to adopt a new analytic architectures and methods

Opportunities for a New Approach to Analytics (Continue

Big Data Ecosystem



Considerations for Big Data Analytics

Criteria for Big Data Projects

1. Speed of decision making
2. Throughput
3. Analysis flexibility

New Analytic Architecture

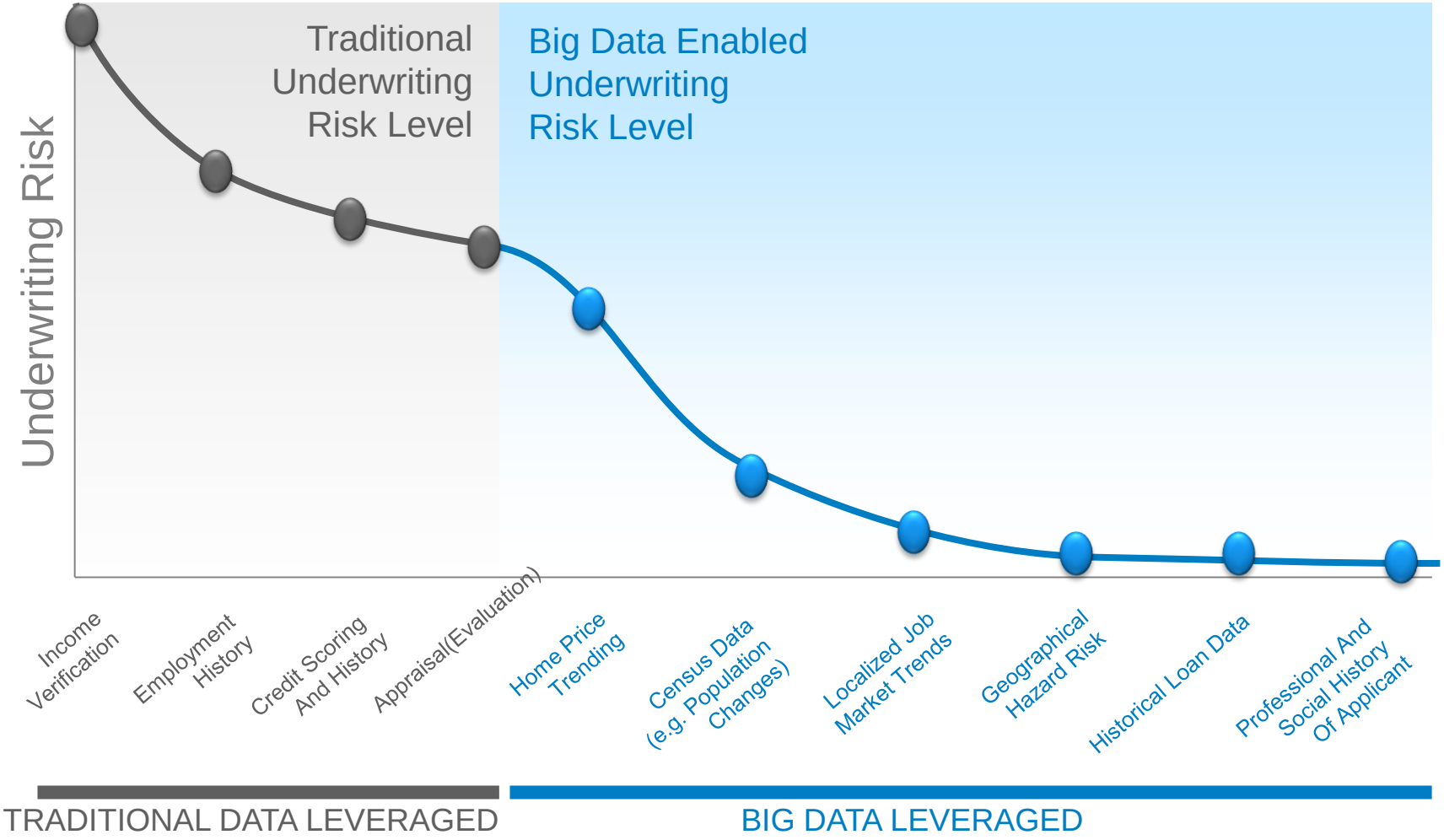
Analytic Sandbox

Data assets gathered from multiple sources and technologies for analysis

- Enables high performance analytics using in-db processing
- Reduces costs associated with data replication into "shadow" file systems
- “Analyst-owned” rather than “DBA owned”

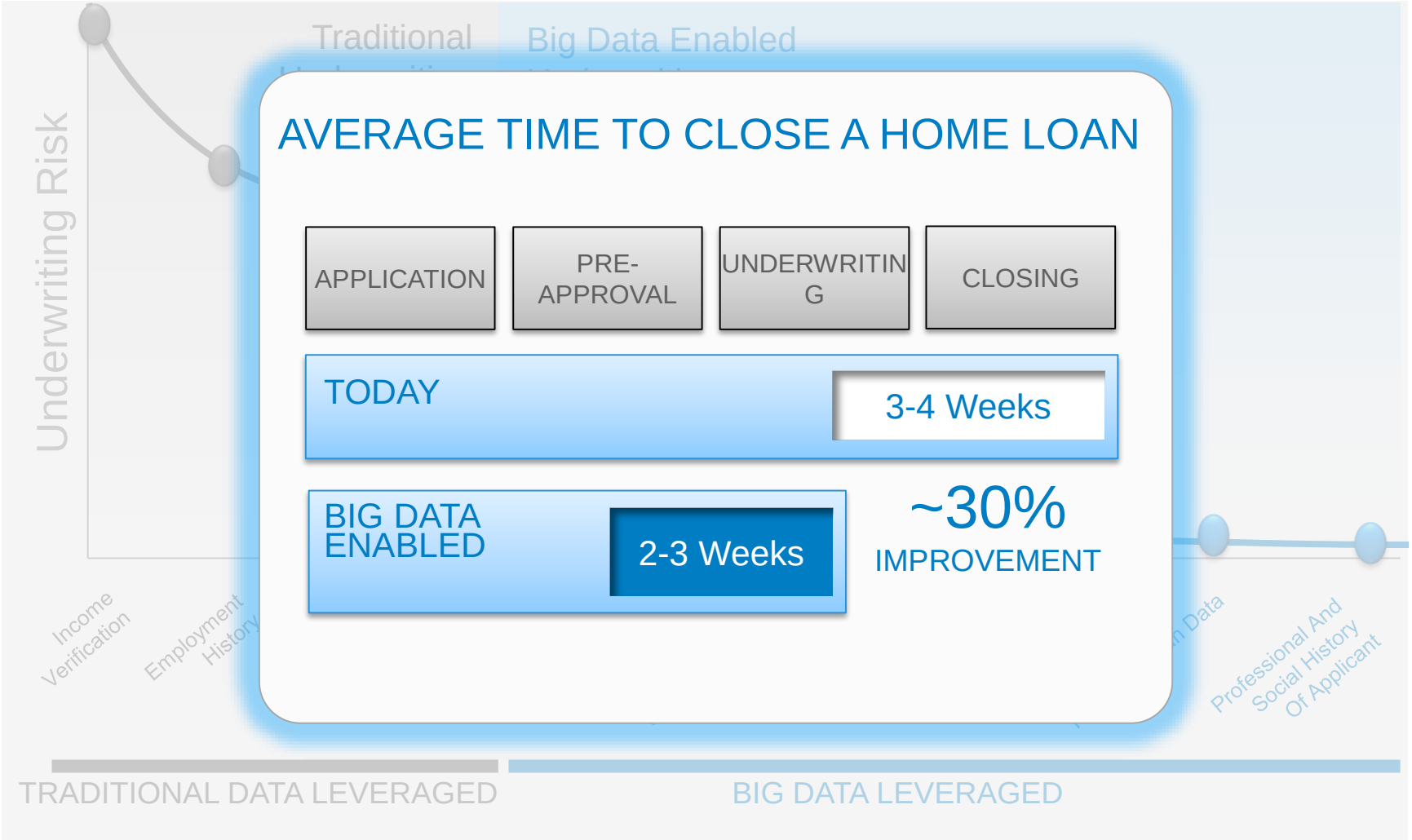
State of the Practice in Analytics: Mini-Case Study

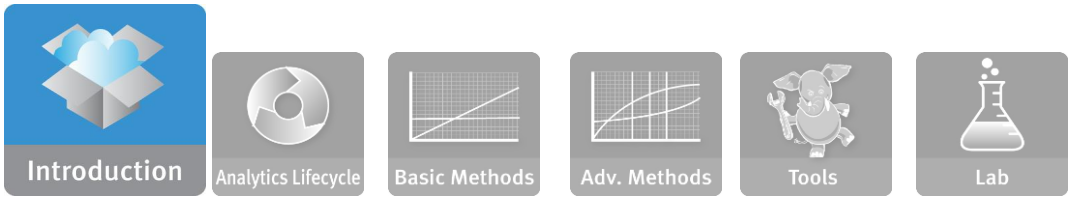
Big Data Enabled Loan Processing at YoyoDyne



State of the Practice in Analytics: Mini-Case Study

Big Data Enabled Loan Processing at YoyoDyne



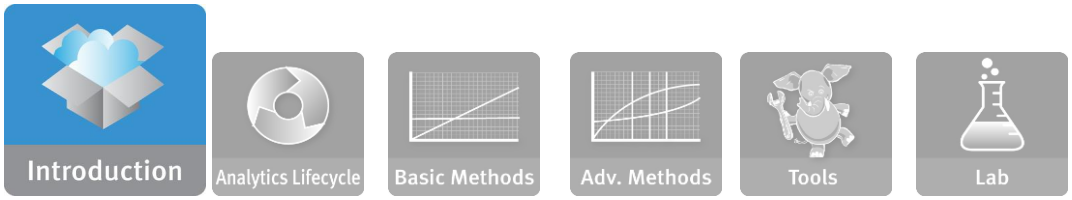


Module 1: Introduction to Big Data Analytics

Summary

During this part the following topics were covered:

- Business drivers for analytics
- Current analytical architecture
- Business intelligence vs. data science
- Drivers of big data and new big data ecosystem



Thanks